

coursera



# The Battle of Neighborhoods : New York

Ashfaql Alam Joarder

Coursera IBM Data Science Capstone Project

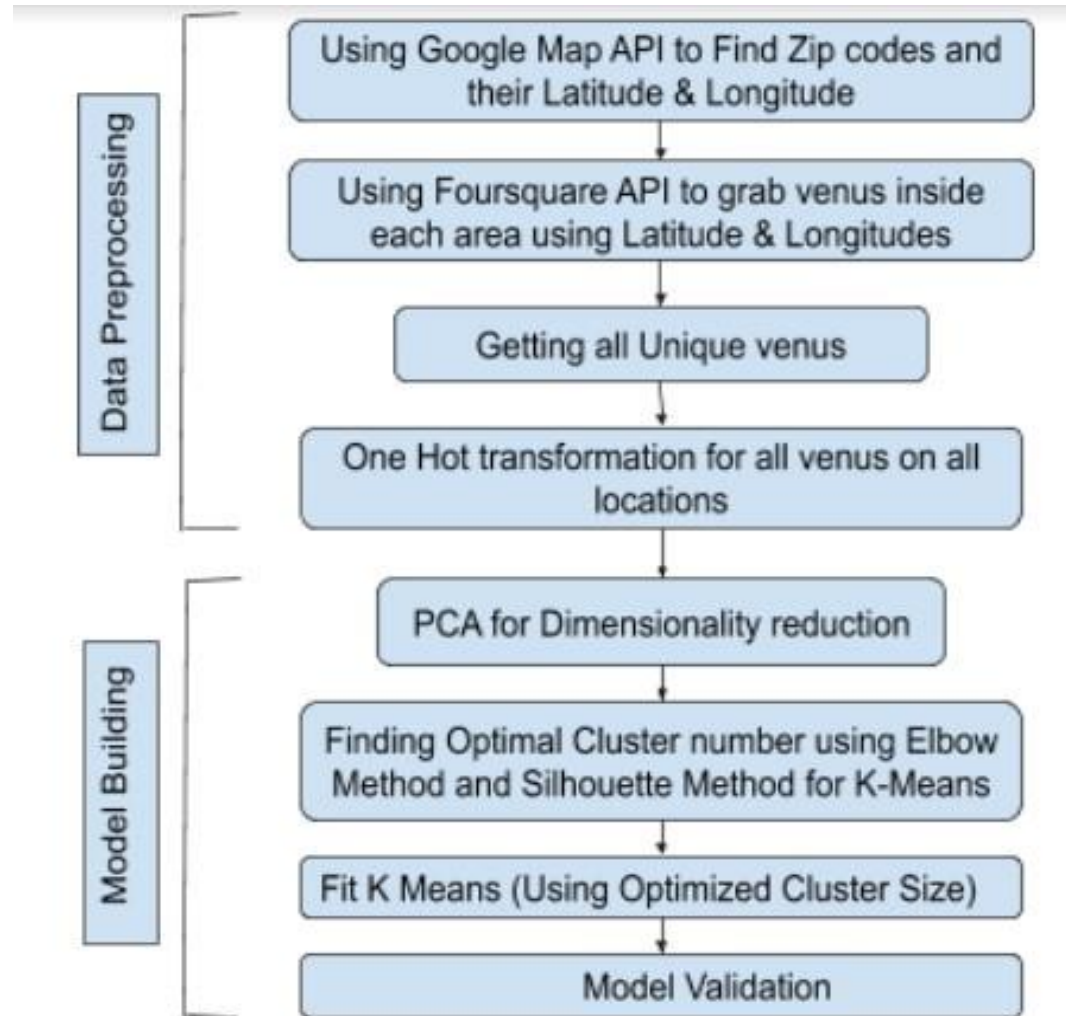
# Motivation

- Suppose, a person has been living in East York, Toronto for 15 sweet years of his/her life.
- Now he has to leave East York and relocate to NY for a change in his job location or some other event.
- Now, he has been used to a particular lifestyle for a longtime. He may likes to go to Indian restaurants for breakfast, maybe he loves to visit some kind of park in the weekends
- Now, he would more like to choose a neighborhood in Manhattan which has all the amenities he was used to in a close proximity.

# Objective

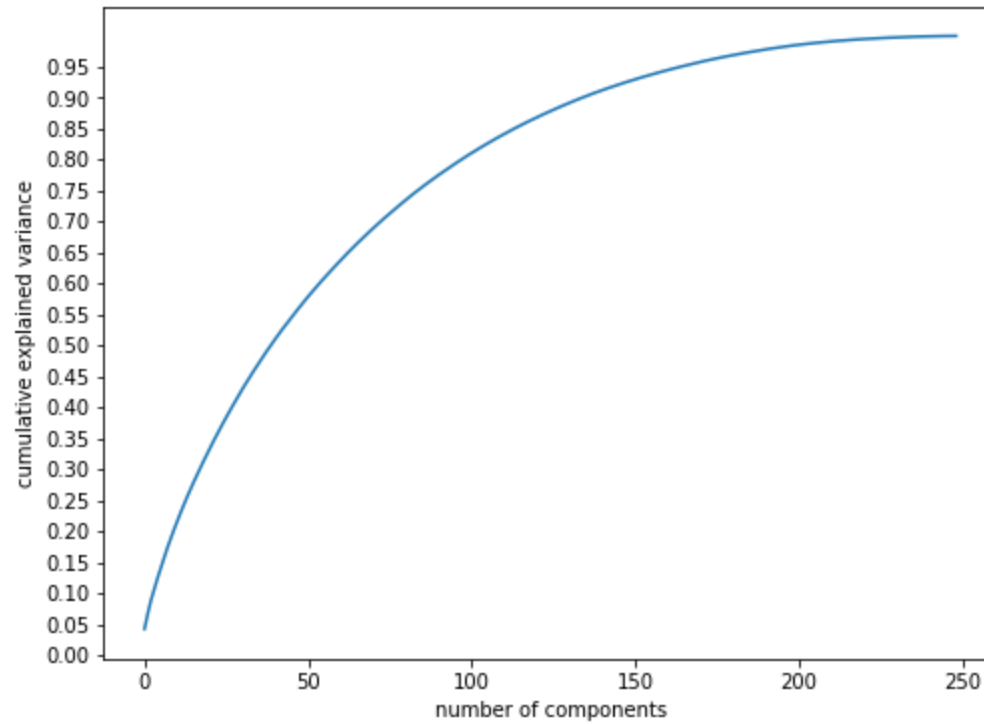
- Applying k-mean clustering algorithm to cluster the neighborhood based on their similarities in different amenities and venues.
- For defining success we will try to figure out the optimal cluster size by doing some exploratory data analysis on different clusters and trying to observe their similarities.

# Workflow



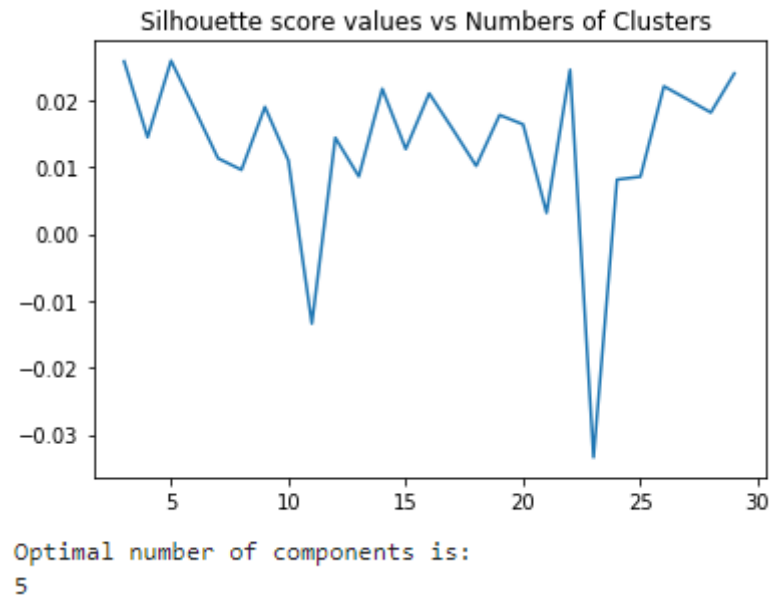
*Figure 1. Neighborhood segmentation work flow chart*

# Selecting Principal Components



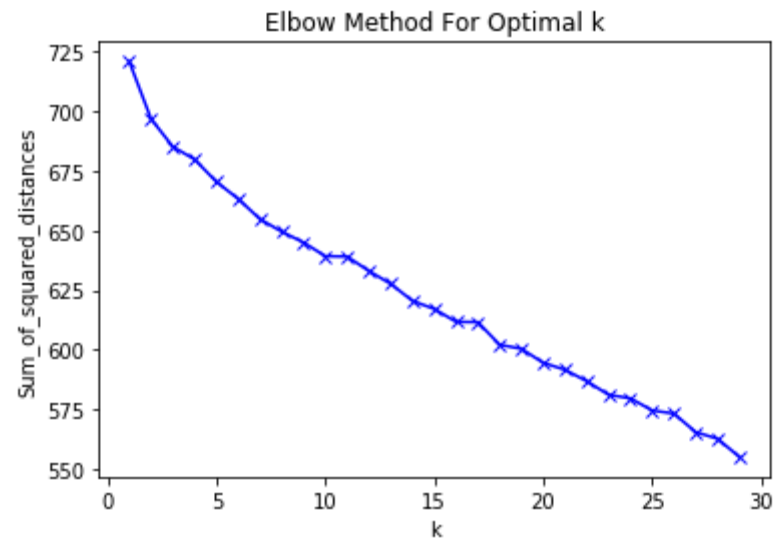
*Figure 2. Selecting Principal Components*

# Silhouette Score



***Figure 3. Silhouette Score confirms optimal Cluster Number 5***

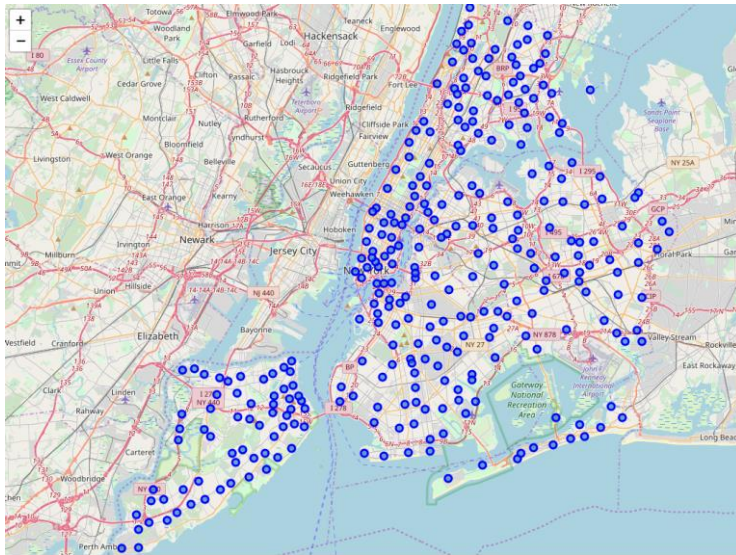
# Elbow Method



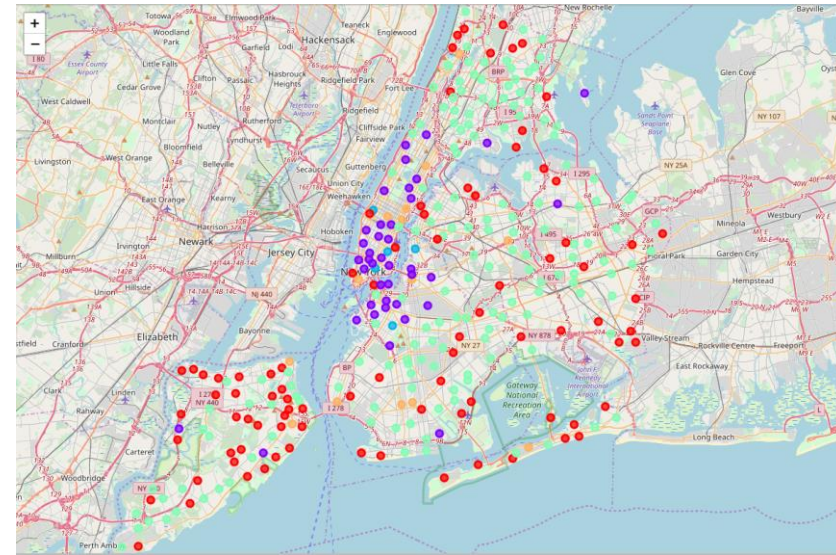
*Figure 4. Elbow found at  $k = 5$  ( $K$  is number of Clusters)*



# Cluster Visualization



*Before Clustering*



*After Clustering*

*Figure 5. NY Zip codes with the assigned cluster label*



# Conclusion

- The project work was only done on the zip codes of New York which includes 401 zip codes each having 150 features even after dimensionality reduction with PCA. The problem is that we have a huge feature space but limited number of samples. We can collect data from entire United States , which will make our dataset well balanced.
- From figure 5 and 6, we can spot certain outlier in our data. In future we will try to filter out those outliers for more robust clustering.
- There could be other clustering algorithms that can work better. In future, DBSCAN seems to be a good fit for our data.
- We can sum everything, and convert to a neighborhood recommendation APP.

**THANK YOU**